Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Presented by:

Samata Parulekar

Contents

- Problem Statement
- Variables in Dataset
- Exploratory Data Analysis(EDA)
- ML algorithms
- Conclusion

Problem Statement:

This dataset consists of tv shows and movies available on Netflix as of 2019.The dataset is collected from Flixable which is a third-party Netflix search engine. In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

In this Project we have done,

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

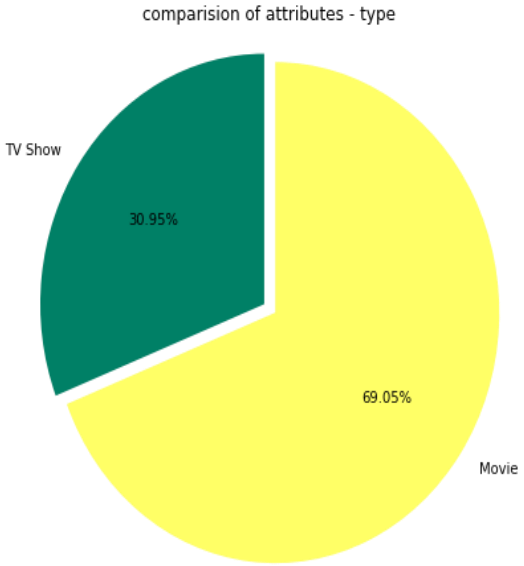
Variables in dataset:

The dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc

- **show_id :** Unique ID for every Movie / Tv Show
- **type :** Identifier - A Movie or TV Show
- **title :** Title of the Movie / Tv Show
- **director :** Director of the Movie
- **cast** : Actors involved in the movie / show
- **country :** Country where the movie / show was produced
- **date_added :** Date it was added on Netflix
- **release_year :** Actual Release Year of the movie / show
- **rating :** TV Rating of the movie / show
- **duration :** Total Duration - in minutes or number of seasons
- **listed_in :** Genre
- **description:** The Summary description

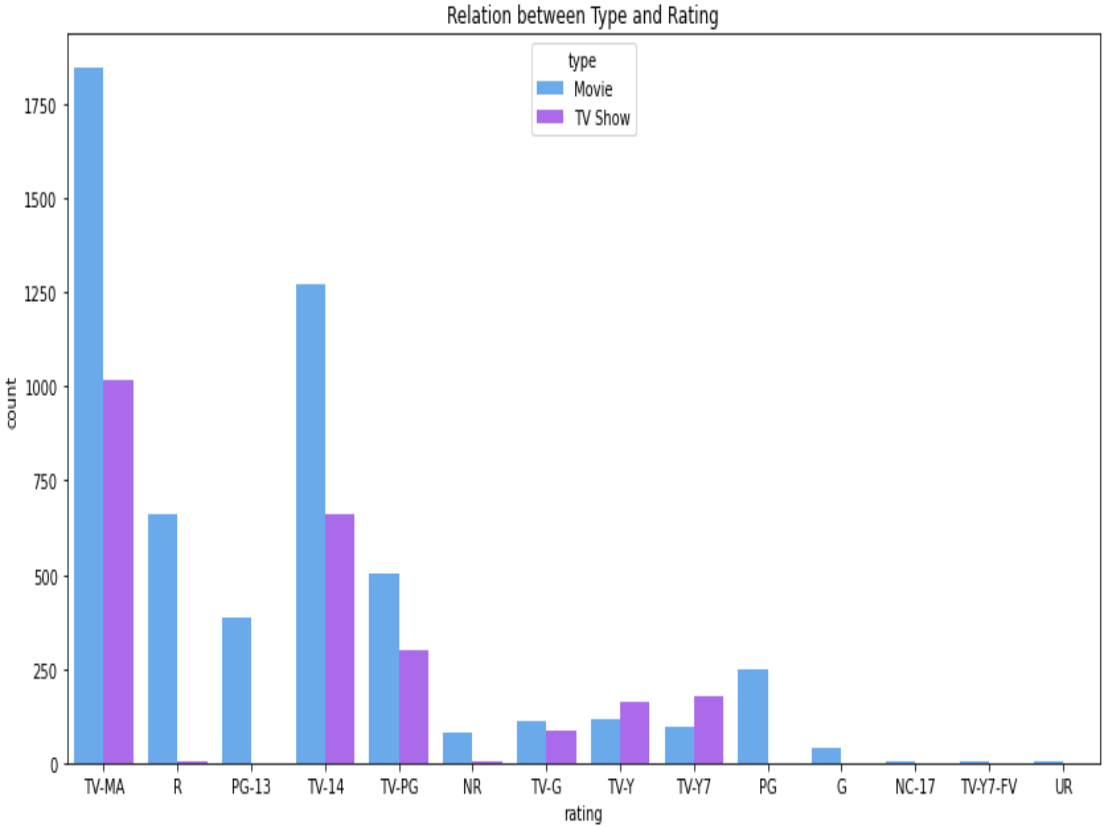
Exploratory Data Analysis (EDA)

1) Which one is most preferred TV shows or Movies to watch?



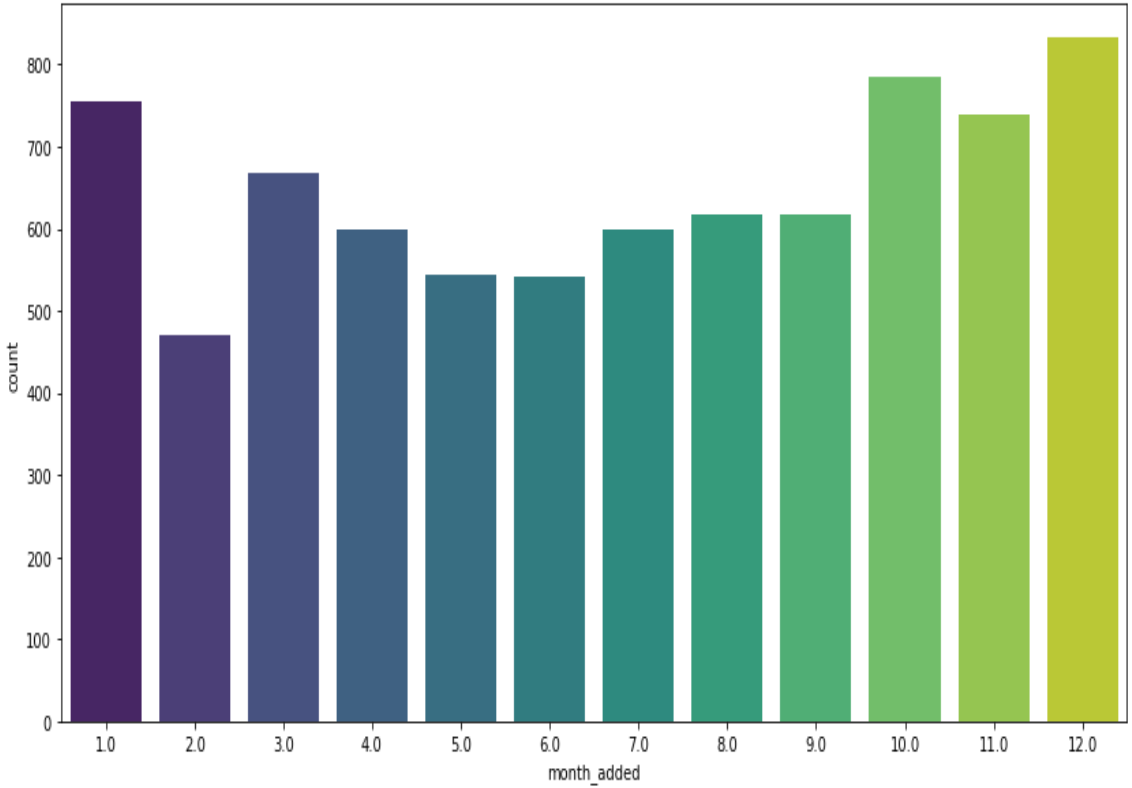
It is clearly visible from the pie-chart that people used to watch movies more than TV shows. Around 69.05% people used to watch movies and 30.95% people used to watch TV shows.

2) Which type of the show(movies or TV shows) got highest ratings?



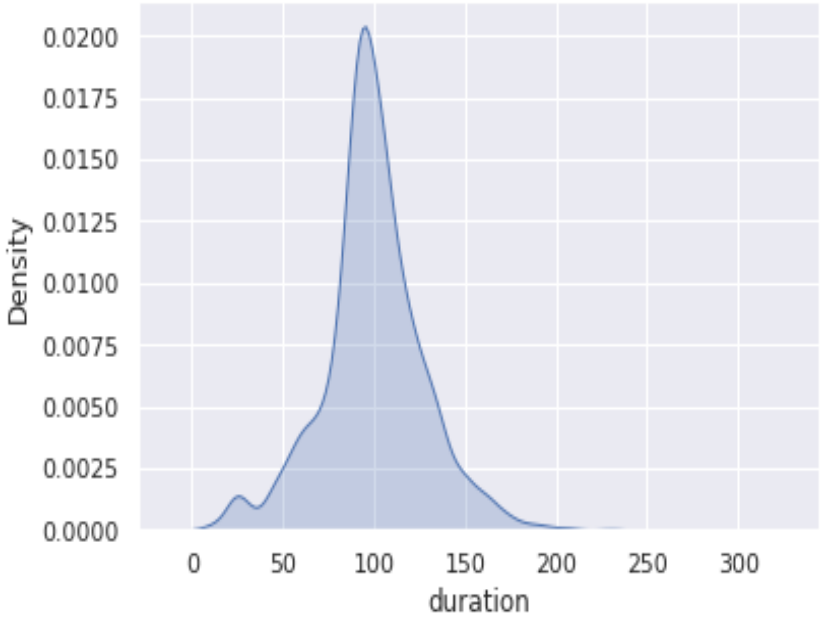
From the bar graph, Movies and TV shows highest ratings given by matured audience only(TV-MA). In that particularly movies got highest ratings as compared to TV shows.

3) which month most of the movies got released?



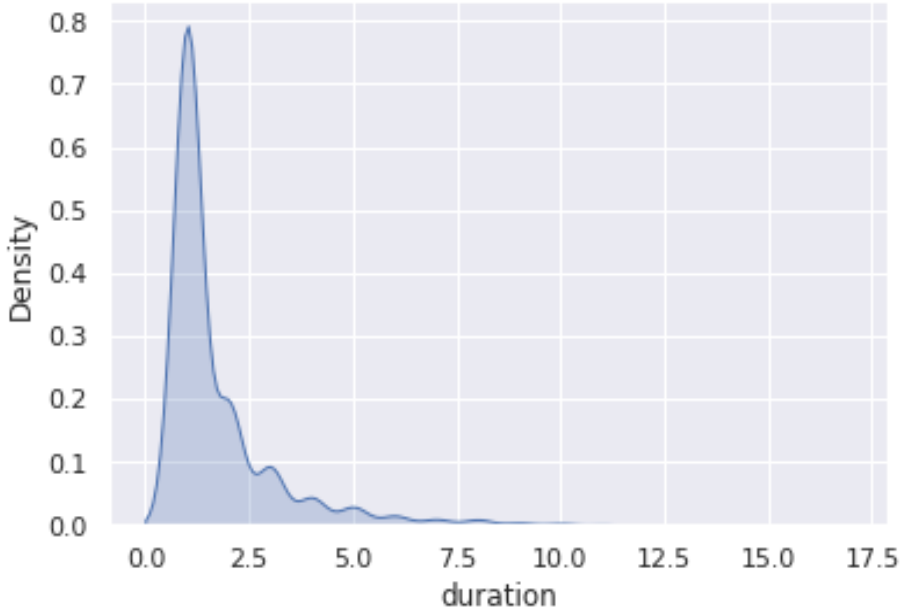
From the above graph, more than 800 movies got released in the month of December and in the month of February the number of movies got released is less.

4) Analysis Of the duration of movie



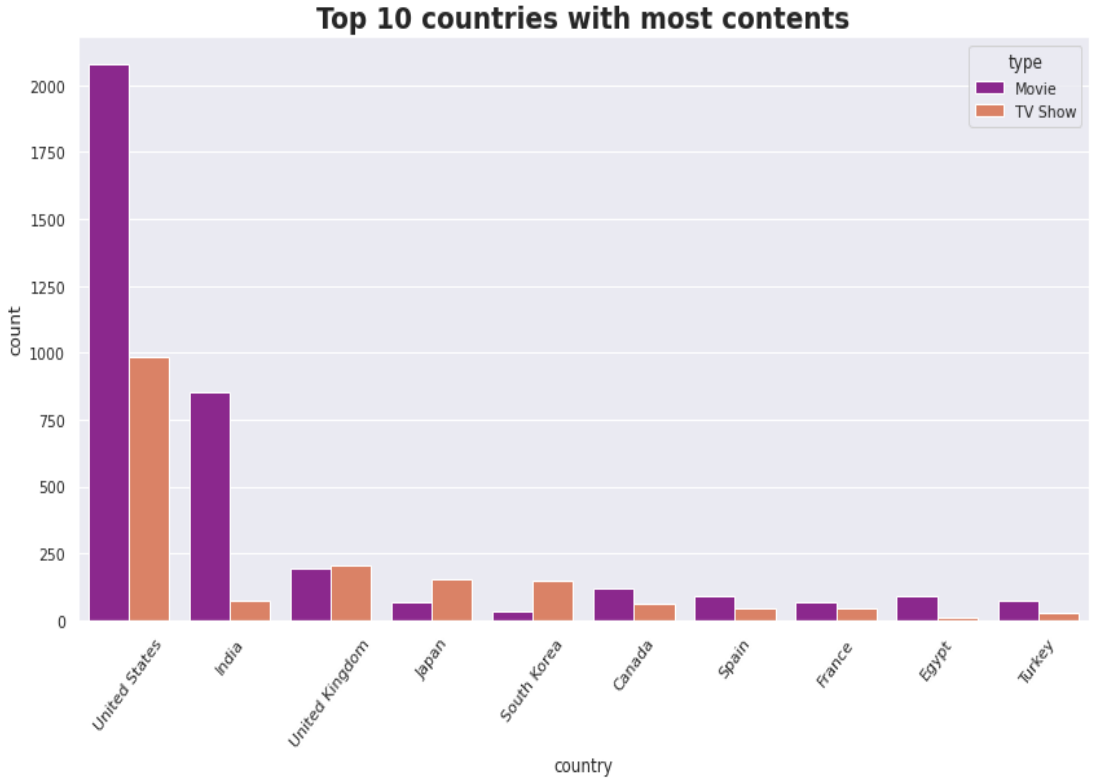
So, a most of the movies are having a duration of 75-120 minutes. This is by taking a fact into an account that people can easily watch 3 hours movie easily.

5) Analysis Of the Number of seasons of TV Shows



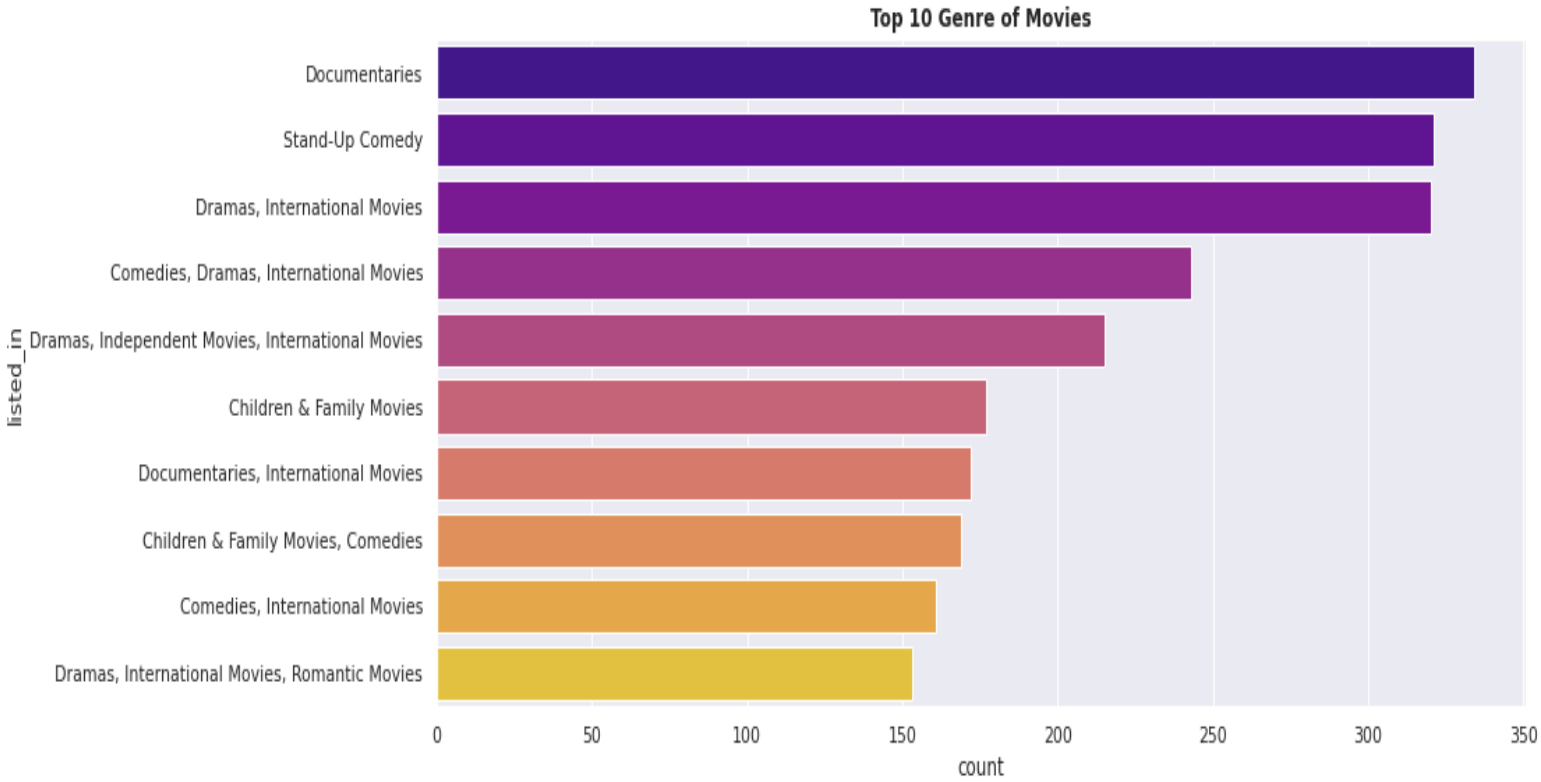
From the graph, most of the TV shows comes up with Two seasons and only few comes up with more than 2 seasons that is purely based on audience response.

6) Which country people used to watch more TV shows and movies?



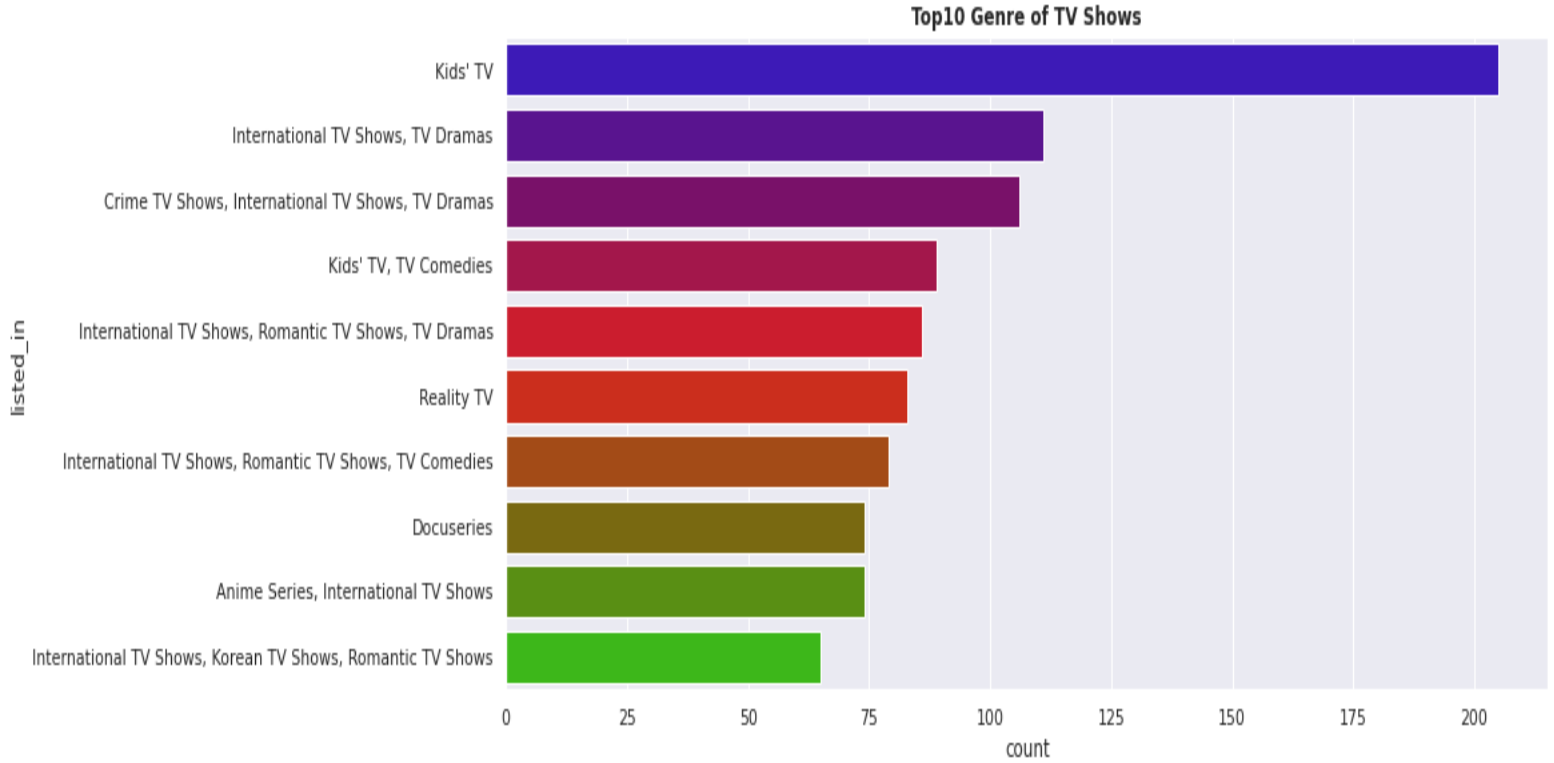
United state people used to watch more movies and TV shows as compared to other countries and more precisely they used to watch more movies than TV shows.

7) Top 10 Genre of Movies



From the above graph Documentaries are the top most genre in Netflix which is followed by stand-up comedy and Drams and international movies

8) Top 10 Genre of TV shows



kids tv is the top most TV show genre in Netflix

ML algorithms(unsupervised)

AI

1) K-Means: K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

a) Sum of squared distance: It is defined as the sum of the squared distance between the average point (called Centroid) and each point of the cluster.

b) Silhouette Coefficient: silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Silhouette score for 15 clusters

For n_clusters = 2, silhouette score is 0.42825796837628916

For n_clusters = 3, silhouette score is 0.3833520787206349

For n_clusters = 4, silhouette score is 0.37412764256018943

For n_clusters = 5, silhouette score is 0.3723122952870676

For n_clusters = 6, silhouette score is 0.36701461874566504

For n_clusters = 7, silhouette score is 0.3761611034643864

For n_clusters = 8, silhouette score is 0.3690808882255994

For n_clusters = 9, silhouette score is 0.37546795077279416

For n_clusters = 10, silhouette score is 0.362320975097743 64

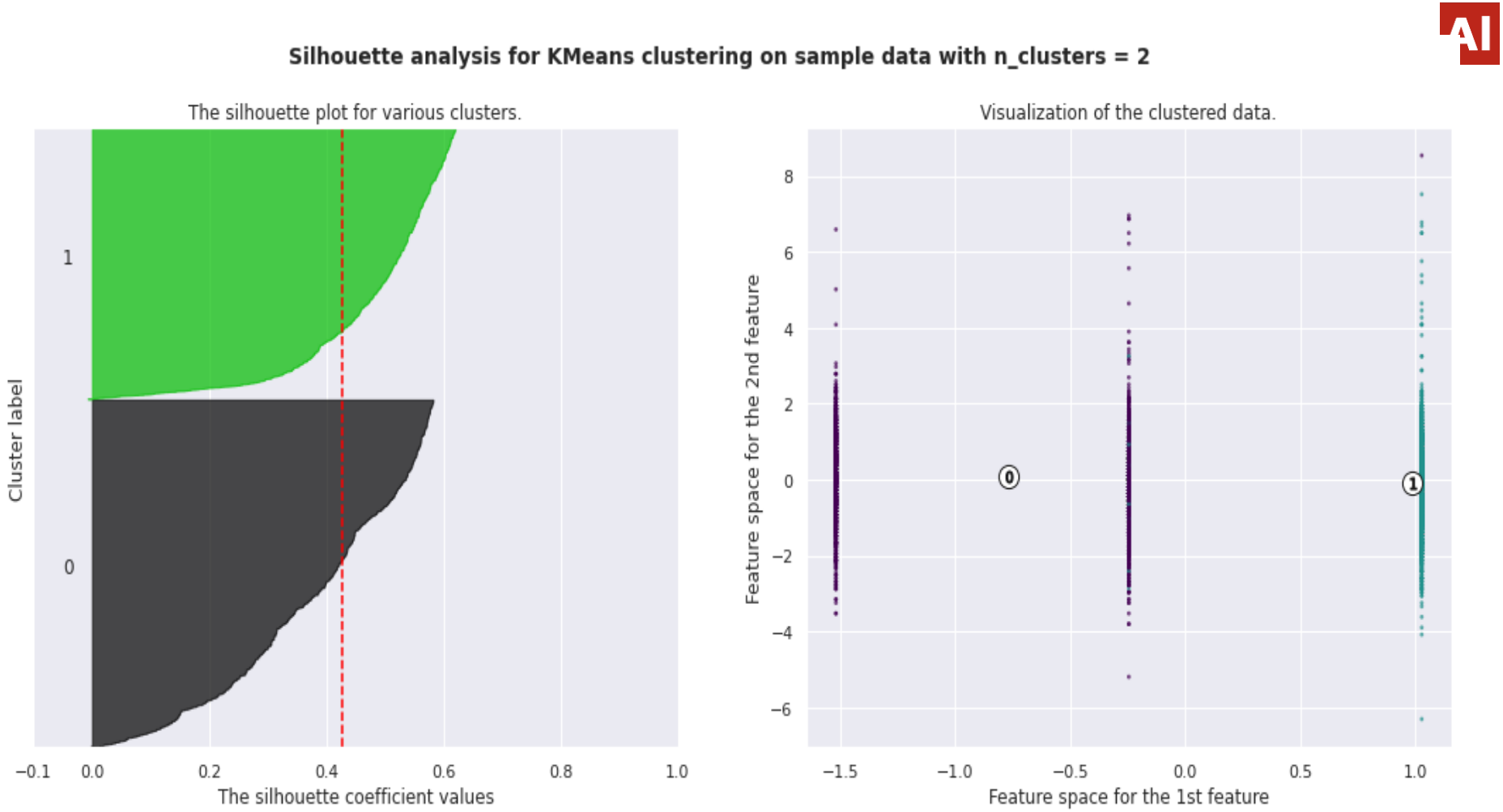
For n_clusters = 11, silhouette score is 0.36173313728750245

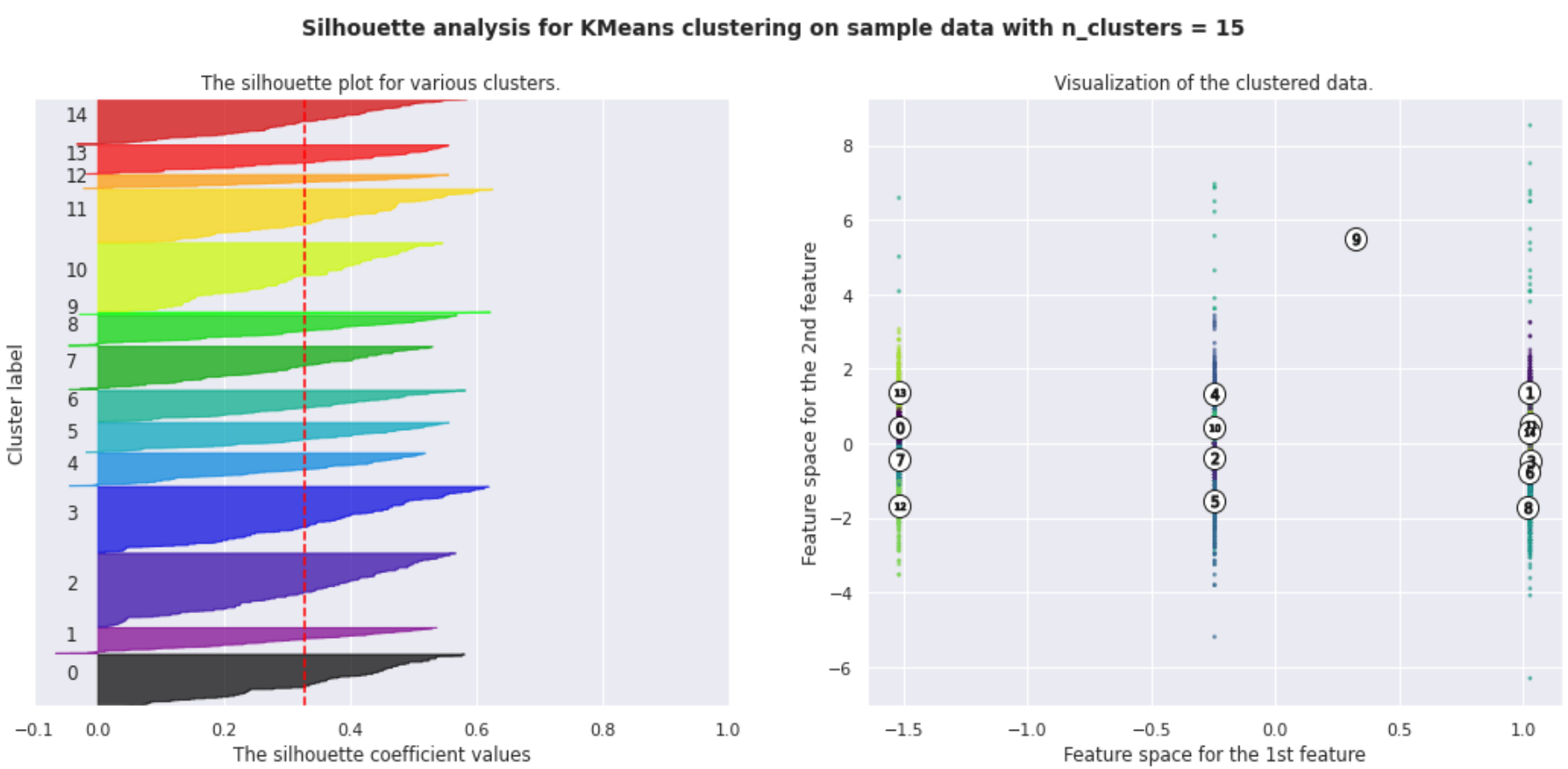
For n_clusters = 12, silhouette score is 0.34973227295318854

For n_clusters = 13, silhouette score is 0.3499832633846903

For n_clusters = 14, silhouette score is 0.3380134011125446

For n_clusters = 15, silhouette score is 0.3296838973160366

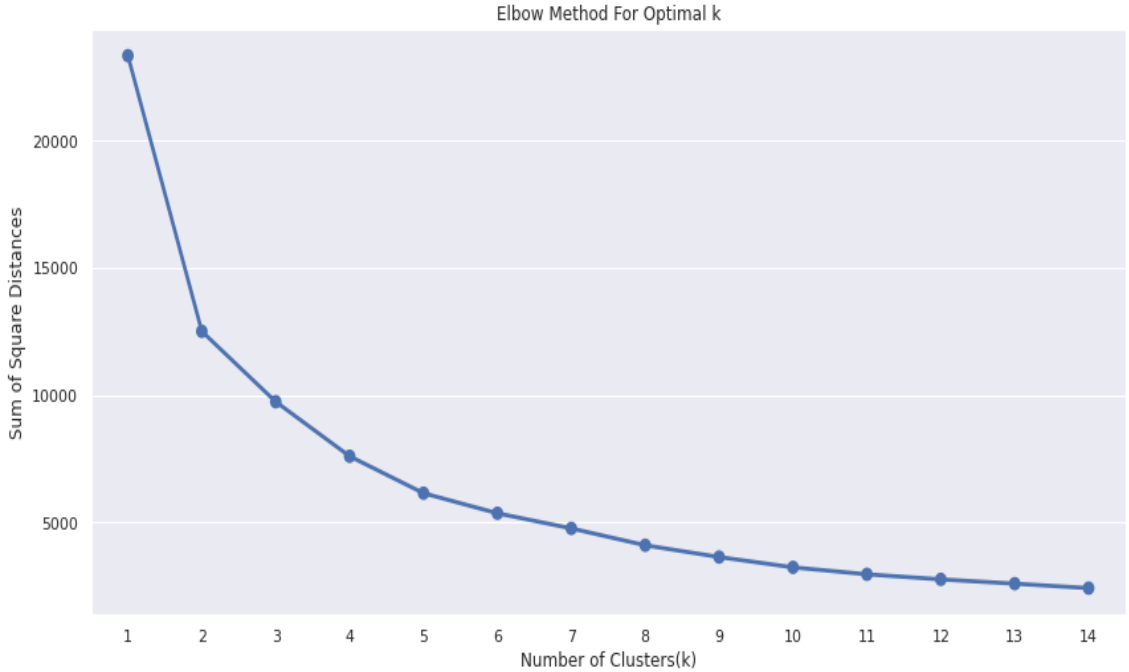

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
Visualization of the clustered data.
Feature space for the 2nd feature
8
6
4
2
0
−2
−4
−6
0
1
−1.5
−1.0
−0.5
0.0
0.5
1.0
Feature space for the 1st feature


Silhouette analysis for KMeans clustering on sample data with n_clusters = 15
The silhouette plot for various clusters.
Cluster label
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
Visualization of the clustered data.
Feature space for the 2nd feature
8
6
4
2
0
−2
−4
−6
−1.5
−1.0
−0.5
0.0
0.5
1.0
Feature space for the 1st feature

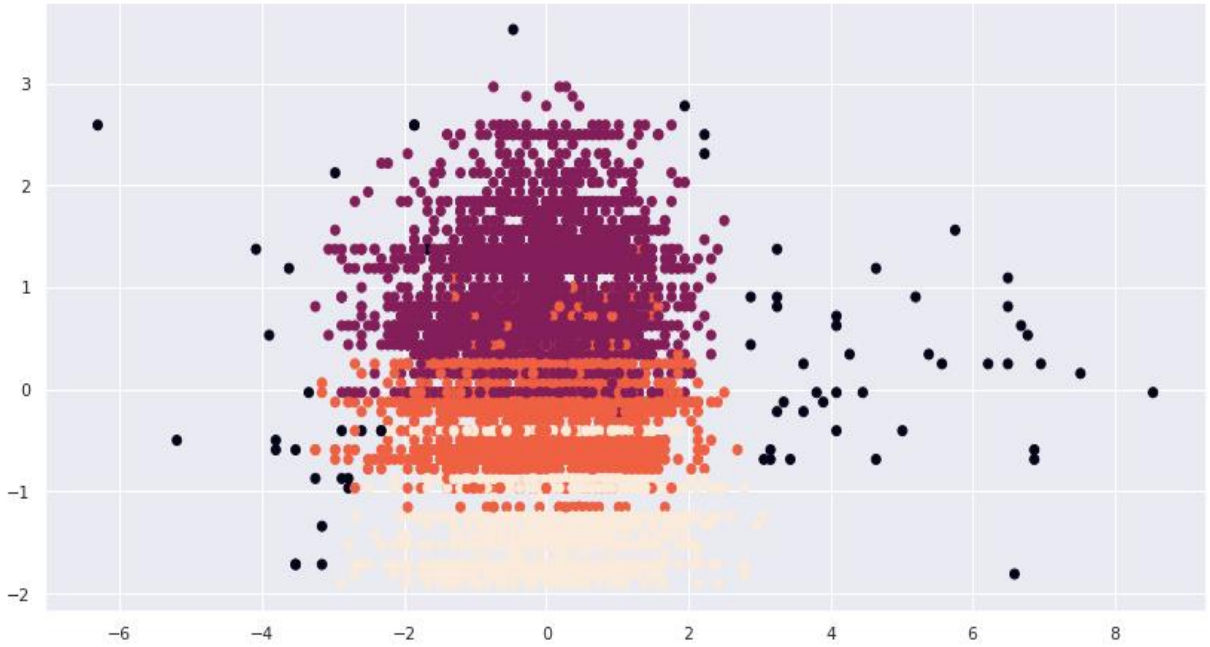
Visualization graphs are plotted for all 15 clusters.

c) Elbow Method: In cluster analysis, the elbow method is **a heuristic used in determining the number of clusters in a data set**. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. This method is used to determine the optimal value of K. Joining point on the elbow curve is the optimum point.

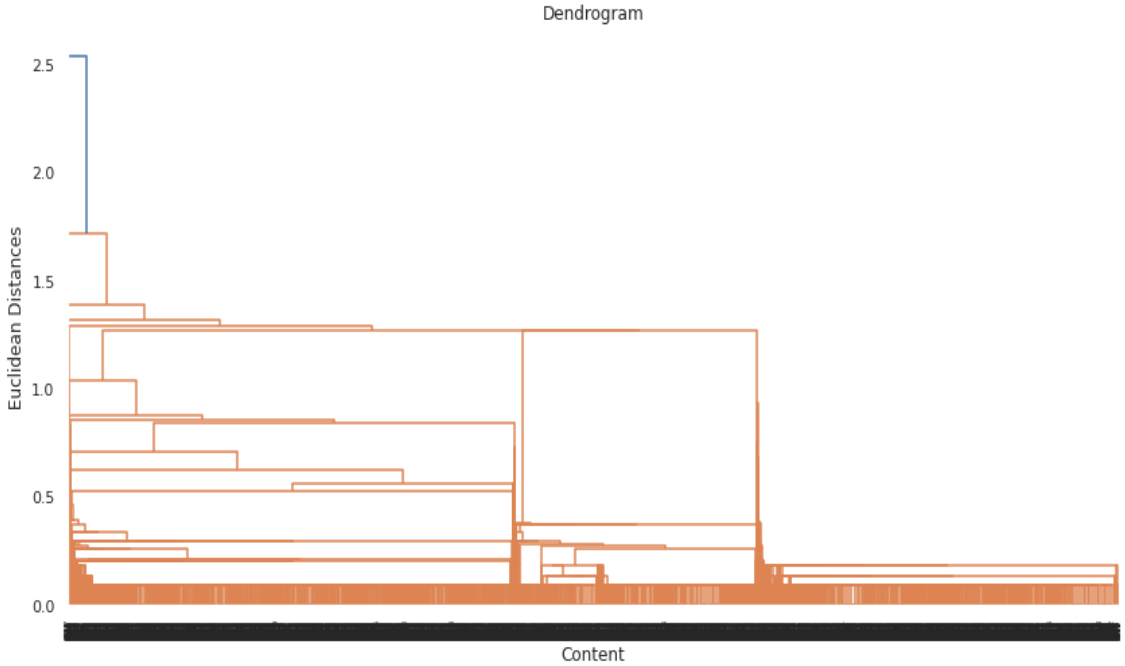
From the above elbow graph K=2 and K=3 are considered as optimum points.



2) DBSCAN Algorithm: It stands for Density-Based Spatial Clustering of Applications with Noise. It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages. In this algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.



3) Agglomerative Hierarchical algorithm: The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure. In agglomerative clustering no need to give the value of k beforehand



Conclusion

- With the help of silhouette score ,optimality test performed for 15 clusters. And we obtained K=2 as a optimal point with the help of elbow method and K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.

- DBSCAN used to show the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.

- Netflix has 5372 movies and 2398 TV shows, there are more movies on Netflix than TV shows. people used to watch movies more than TV shows. Around 69.05% people used to watch movies and 30.95% people used to watch TV shows.

- Movies and TV shows highest ratings given by matured audience only(TV-MA). In that particularly movies got highest ratings as compared to TV shows.

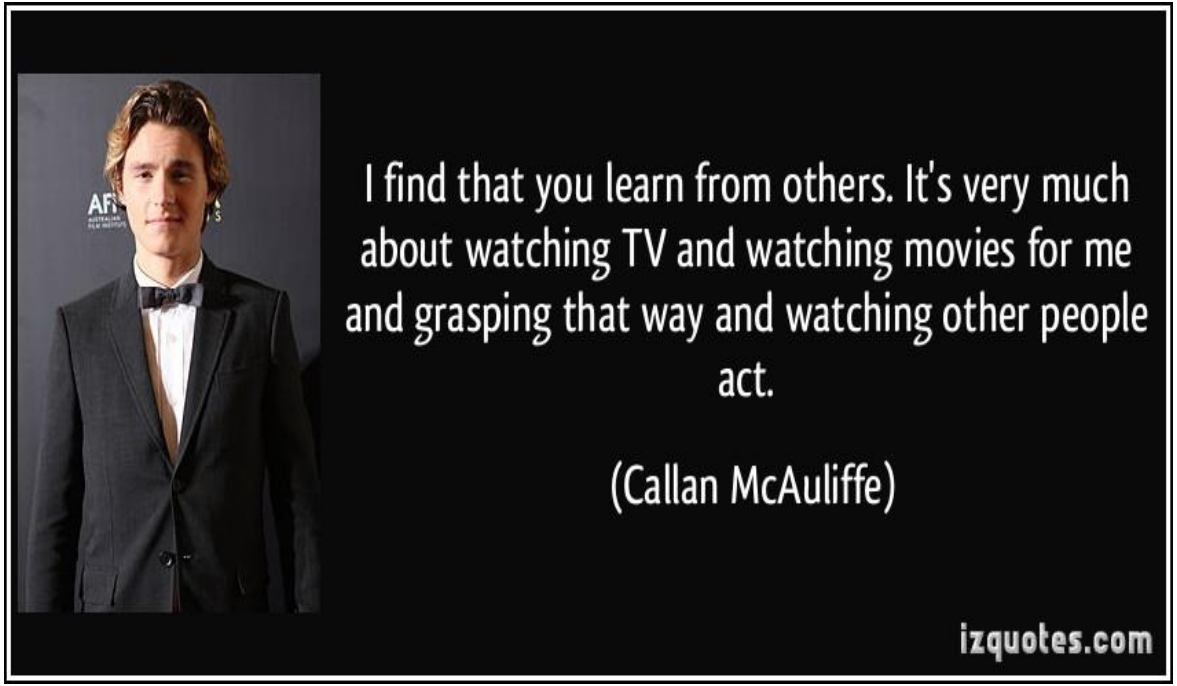
- More than 800 movies got released in the month of December and in the month of February the number of movies got released is less.

- Most of the movies are having a duration of 75-120 minutes. This is by taking a fact into an account that people can easily watch 3 hours movies and most of the TV shows comes up with Two seasons and only few comes up with more than 2 seasons that is purely based on audience response.

- United state people used to watch more movies and TV shows as compared to other countries and more precisely they used to watch more movies than TV shows.

- Documentaries are the top most genre in Netflix which is followed by stand-up comedy and Drams and international movies and kids tv is the top most TV show genre in Netflix

THANK YOU.....


I find that you learn from others. It's very much about watching TV and watching movies for me and grasping that way and watching other people act.
(Callan McAuliffe)
izquotes.com